What are the “phonemes” in phoneme-grapheme mappings? A perspective on the use of databases for lexicon development

Article (Accepted Version)

Cahill, Lynne (2017) What are the “phonemes” in phoneme-grapheme mappings? A perspective on the use of databases for lexicon development. *Journal of Written Language and Literacy*, 20 (1). pp. 104-127. ISSN 1387-6732

This version is available from Sussex Research Online: <http://sro.sussex.ac.uk/id/eprint/74183/>

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:

Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

What are the “phonemes” in phoneme-grapheme mappings? A perspective on the use of databases for lexicon development

Lynne Cahill, University of Sussex

Abstract

The CELEX lexical database (Baayen, Piepenbrock & van Rijn 1995) was developed in the 1990s, providing a database of the syntactic, morphological, phonological and orthographic forms of between 50,000 and 125,000 words of Dutch, English and German. This database was used as the basis for the development of the PolyLex lexicons, which included syntactic, morphological and phonological information for around 3,000 words of Dutch, English and German. Orthographic information was subsequently added in the PolyOrth project. The PolyOrth project was based on the assumption that the underlying, lexical phonological forms could be used to derive the surface orthographic forms by means of a combination of phoneme-grapheme mappings and sets of autonomous spelling rules for each language. One of the complications encountered during the project was the fact that the phonological forms in CELEX were not always genuinely underlying forms which made deriving the orthographic form tricky. This paper discusses the nature of lexical databases and the implications for using such databases for research on writing systems and their relationship to phonology.

Keywords

phoneme-grapheme mappings, lexical databases, lexicons, underlying phonology, lexical phonology post-lexical phonology

1 Introduction

The relationship between spoken language and alphabetic scripts is often characterised, at least in part, in terms of the mappings between phonemes and graphemes. Aspects of both phoneme-to-grapheme and grapheme-to-phoneme mappings including their definition, deduction and learning, are fundamental to linguistic research into writing systems and orthographies, speech analysis and synthesis research, and educational approaches to reading. Among the key questions addressed are those relating to the variable mappings, especially in languages like English which exhibit significant irregularity in the spelling, and issues relating to the units of analysis, where more than one

grapheme or phoneme may be relevant to the mapping (for instance where the spelling of a consonant is dependent on the nature of the vowel preceding or following it). However, the question of the exact status of the “phonemes” in these mappings is rarely addressed. There is an assumption in much of the work in these fields that the phonological representations are well established. However, as this paper shows, the phonological level of representation is actually the part of the equation that is **not** well established, in contrast to the orthographic level, for which there are clearly defined concrete forms.

In order to research the mappings, various kinds of data can be used. For speech research, in particular, raw speech in the form of audio recordings or acoustic signals is the extreme end of the process, either at the beginning or the end, so the intermediate level of “phonemes” is not necessarily important. At least, it is not necessary to have agreed or standardised phoneme representations because different systems simply have to work on the speech signal. For literacy education and writing systems research, however, it is necessary to have an agreed way of talking about pronunciation. In order to comprehensively research a writing system it is vital to have large quantities of representative examples of the (standard) pronunciation and spelling of the words of a language. This information might be provided by dictionaries, which tend to have standard pronunciations and spellings, together with information about definitions and etymology. Dictionaries, however, may not have this information in a format that is both consistent and easily accessible. They also tend to be different for different languages. That brings us to the possibility of using lexical databases for the research. The increased availability of large scale databases of lexical information about a range of languages should be very useful for this kind of research. If we have databases with information about the pronunciation and spelling of large numbers of words then we can use this information to research the relationship between the “phonemes” and the graphemes.

Large-scale lexicons are vital parts of natural language processing (NLP) systems. Speech recognition and generation do not always make use of large lexicons, but many systems do. Developing lexicons for use in such systems may benefit greatly from the use of databases of lexical information, but the lexicons themselves may have specific requirements that mean that they need to be structured differently from (simple) databases. Development of lexicons can also be a means to test theories about the relationships between different kinds of information. The distinction between databases and lexicons is discussed further in subsections 2.2 and 2.3.

The crux of the problem discussed in this paper can be illustrated by the diagram in Figure 1 (from Nunn 1998: 31).

Figure 1: Nunn's two-step model

Nunn's assumption, and one which is commonly proposed for alphabetic systems, is that a common underlying representation of the phonology (the "abstract sounds") of a language can be mapped to both the surface pronunciation and to the spelling. Thus, we should be able to define lexicons with the underlying phonological forms together with one set of rules to derive the surface pronunciation and two sets of rules (phoneme-grapheme conversion rules and autonomous spelling rules) to derive the spelling. The problem, as shall be shown in detail, is that the exact status and nature of the "abstract sounds" in Nunn's model are unclear.

The CELEX lexical databases provide a range of lexical information, including pronunciation and spelling, for Dutch, English and German. They appear to offer a useful resource for research on the relationship between phonemes and graphemes for these three languages. The PolyLex and PolyOrth projects developed highly structured inheritance based lexicons of Dutch, English and German populated with lexical data from the CELEX databases. Three specific relationships are discussed here. The appearance of schwa in English is of interest because the schwa occurs where an underlying full vowel is in an unstressed syllable. It is a case of post-lexical phonology, and in terms of Nunn's model we would expect the underlying vowel to be the full vowel, in order to map to the correct vowel in the spelling. Final consonant devoicing is similarly a post-lexical process, affecting obstruent consonants in coda position. In the majority of cases, the spelling reflects the underlying consonant before devoicing has taken place, although there are examples in Dutch where this is not the case. Ablaut and umlaut, on the other hand, are lexical processes and are reflected in the spelling. We would therefore expect the underlying forms in Nunn's model to reflect the forms after the ablaut or umlaut processes have taken place.

This paper explores the difficulties that arise when assumptions about the phonological representations in databases turn out to be wrong. Specifically, the use of the CELEX databases to

develop and populate the PolyLex/Orth lexicons resulted in difficulties that were caused by the fact that the phonological representations did not consistently reflect a lexical, underlying form, which is what is required to accurately determine the mappings to graphemes. The next section reviews previous work on phoneme-grapheme mappings, looks at a range of lexical databases and outlines different approaches to lexicon development. The next two sections give detailed descriptions of the CELEX databases and the PolyLex/Orth lexicons, respectively. They are followed by a detailed discussion of three specific (morpho)phonological phenomena that illustrate the issues: the problem of schwa, especially in English; umlaut/ablaut in all three languages; and final consonant devoicing in German and Dutch. Finally, there follows a discussion of the implications of these issues and where researchers wanting to investigate phoneme-grapheme mappings might go for their data.

2 Background

2.1 Phoneme-grapheme mappings

Work on phoneme-grapheme mappings is a key element of research within the speech processing community (see Benetsy, Sondhi & Huang (2008) for a collection of work on speech processing). The direction of the mapping makes a big difference to the approach and the results. When going from graphemes to phonemes and ultimately to speech it is necessary to consider allophonic variation, word boundary phenomena and fast speech processes, and there are, of course, many issues around prosody in producing natural sounding speech. When going from speech to text, the speech stream has to be segmented, at least into words and potentially into segments. Although modern approaches to speech recognition and generation tend not to make use of a specific level of phonemic representation, early work in speech recognition focused on essentially a two-stage process of recognising phonemes within the speech stream and then mapping those phonemes to graphemes to produce a text version of the spoken language.

Nunn's (1998) work on Dutch was one of the inspirations behind the approach to representing orthography within the PolyOrth lexicons. It focuses on Dutch, but the framework proposed should be applicable to any alphabetic writing system. As we shall see below, Nunn discusses some issues surrounding the precise nature of the underlying phonological forms. For example, she proposes that schwa in Dutch does not have an underlying representation at all and she also addresses the question of devoicing, with a handful of examples of Dutch words where the normal approach of the spelling reflecting the voiced variants is not observed. This is discussed in more detail in section 5 below.

Sproat (2000), focusing mainly on English, discusses the question of whether different phonological levels ("deeper" or more "surface" in generative terms) are represented in the orthography.

Specifically he states that the level may be different for different writing systems (hence the notion of deep and shallow orthographies) but that it is consistent for all vocabulary in any individual language. He calls the phonological level represented in phoneme-grapheme mappings the Orthographically Relevant Level (ORL). This raises questions about how, or indeed whether, it is possible to provide databases for different languages that include consistent phonological forms across the languages but which can, nevertheless, be used to determine the relationship between the phonology and the spelling. Herring (2006) proposed that spelling should be viewed as resulting from a combination of phonological and morphological information. The interaction of phonology and morphology can be seen as relevant in Sproat's distinction as well, and we shall see below that the relationship between underlying and surface phonological forms needs to be viewed in the context of the relationship between lemmas and word forms. When we come to the question of what forms should be (and are) included in lexical databases, this relationship will become vital. Figure 2 illustrates the relationships between morphologically underlying and surface (lemma/word form) and phonologically underlying and surface forms. The involvement of morphology in the relationship between pronunciation and spelling is clearly important, and so assuming that we can represent the relationship between pronunciation and spelling without reference to morphology is unlikely to be successful, at least for languages with deep orthographies.

2.2 Databases

For the purposes of this paper we are interested in databases that provide both orthographic and phonological information for the words of one or more languages. Nerbonne (1998) defines linguistic databases as a collection of electronic records of linguistic data. He further stipulates that databases should be declarative, that is, not be restricted to any particular software package, and that they should be consistent. When looking for databases of both phonological and orthographic information, we might consider standard dictionaries as providing this. However, even electronic versions of dictionaries tend not to fulfil both of the above requirements of declarativeness and consistency. Most machine readable dictionaries are accessed via graphical user interfaces and few provide exhaustive consistent representations of both the orthographic and phonological forms of all word forms¹, relying, for example, on abbreviations to encode regular inflections.

There are a number of databases that do fulfil Nerbonne's criteria. The LEXIQUE lexical databases of French (New, Pallier, Brysbaert & Ferrand 2004) contain 135,000 word forms from 55,000 lemmas of French. The information provided includes (morpho)syntactic and frequency information, as well as phonological information which includes a basic phonological representation, number of phonemes/letters, syllable structure and "abstract orthographic and phonological structure". Both abstract structures in fact consist of CV-templates of each orthographic and phonological form, so the word *chat*, /ʃa/ has the abstract orthographic representation CCVC, and the abstract phonological representation CV. The phonological representations were derived from a text-to-speech application, with manual correction to iron out exceptional pronunciations, eliminate the distinction between different pronunciations of /a/, /o/ and /r/ and "correct the phonetic representation for the schwa" (New et al. 2004: 517).

The CLEARPOND database (see Marian, this issue) includes orthographic and phonological neighbourhood frequencies for five languages: English, French, Spanish, Dutch and German. That is, a word like *tone* has orthographic neighbours *bone*, *ton*, *one* etc.. It also has phonological neighbours *bone*, *toe*, *tow*, *loan* etc.. The databases are derived from subtitle corpora in which film and television spoken utterances correspond with written subtitles. The phonological forms are important within CLEARPOND for measuring phonological neighbourhoods. When it comes to questions such as whether a vowel should be represented with a schwa or a full vowel, this is not likely to have a very significant impact on the neighbourhood measurements, so the fact that a genuinely underlying phonological representation is not used in these databases is not a problem.

¹ "Word form" is used throughout, even when referring to the CELEX databases, for which the term is written as "wordform" in the documentation.

The CELEX lexical databases include information about Dutch, English and German. The information included is syntactic, morphological, phonological and orthographic. The databases are intended to be general purpose repositories of lexical information, represented consistently across the three languages. The CELEX lexical databases were chosen as the basis for development of the PolyLex/PolyOrth lexicons (see below) because of the apparent comprehensiveness and consistency of coverage of phonology and orthography for word forms and lemmas across all three languages. These databases are described in detail in section 3 below.

2.3 Lexicons

If databases are collections of linguistic data that are declarative and consistent, lexicons need not necessarily be either. Lexicons for NLP applications need to serve the purpose of the application in question. This may mean simply a list of words with grammatical category (e.g. Carroll & Grover 1989) or it may mean a highly structured system for matching words with a variety of lexical information including syntax, morphology, semantics and even pragmatic or discourse information (e.g. Evans, Piwek, Cahill & Tipper 2008). Another possible purpose of lexicons is to test theories of lexical representation. That is, (computational) lexicons might be developed in order to provide testable models of linguistic or psycholinguistic theories of the lexical relationships within a language (e.g. Network Morphology (Brown & Hippiusley 2012)). The lexicons focused on in this paper have the latter purpose.

The PolyLex lexicons of Dutch, English and German (Cahill & Gazdar 1999) were developed to test theories of the interaction of phonology and morphology and how this interaction can be represented in a hierarchical lexicon. The project also focused on multilingual questions of how common information can be shared, especially across closely related languages, but those questions are not of interest for our present purposes.

The PolyOrth lexicons (Cahill, Herring & Tiberius 2013) developed the PolyLex lexicons further to include orthographic information and to address the relationship between orthography, phonology and morphology. These lexicons are discussed in detail in section 4 below.

3 CELEX

The CELEX lexical databases were developed in the 1990s at the Centre for Lexical Information, Nijmegen. They include databases Dutch, English and German. For each language in release 1 there is a total of nine databases, representing different kinds of information about word forms and lemmas. For lemmas there are five databases: orthography, phonology, morphology, frequency and syntax. For word forms, only the first four of these are provided (since syntactic information is only

applicable to lemmas). The number of lemmas and word forms varies across the languages. Table 1 shows the figures for the first release of CELEX.

	Dutch	English	German
Lemmas	124,000	52,500	50,000
Word forms	381,000	160,000	360,000

Table 1: numbers of lemmas and word forms in CELEX

These figures, in themselves, tell us something about the morphological complexity of the three languages. For example, we can see from these that for every lemma in English and Dutch, there are on average around three word forms, while for German the average is over seven. On the other hand, the difference in the number of lemmas for each language does not reflect any significant linguistic differences – they are simply the result of different amounts of effort by different groups working on the three languages. The databases that we are interested in here are the orthography and phonology databases.

In release 2, for English and Dutch but not German, databases of syllable frequency were included. These databases are interesting for us because they are described in the README files thus:

“The dfs.cd file lists **phonetic** syllable frequencies calculated from a combination of the **orthographic** word form frequencies and the syllabified **phonemic** word form transcriptions in the CELEX database.”² (my emphasis)

The reference here to “phonemic ... transcriptions” is interesting. The word “phonemic” is not used elsewhere in the CELEX documentation and, as we shall see, it is clearly not an accurate description of the type of transcription actually provided.

The CELEX databases are provided in basic ASCII format, with records for each lemma and word form comprised of a number of distinct fields of information. CELEX comes with the FLEX software, which provides tools to extract lexicons from the databases. This reflects Nerbonne’s assertion that databases should be declarative, the ASCII format being fully portable, in contrast to the derived lexicons.

² From the README file in CELEX/DUTCH/DFS on release 2 CELEX CD.

3.1 Lemmas and word forms

On one level, chiefly morphological, it is clear what the relationship between lemmas and word forms is. Each lemma is equivalent to a lexical or dictionary entry, which can be realised as a number of different word forms depending on the morphological form and therefore the inflections. However, when we come to think about the exact status of the phonological and orthographic forms given at the lemma and word form levels it becomes clear that there is a more complex relationship, and one that is fundamental to the discussion here.

If a lemma is not an actual word form then what exactly is represented by the lemma phonology? We might expect the phonology associated with a lemma to be an underlying representation of its root (or stem). That would mean that we would expect the phonological form of a lemma to be the form before either lexical or post-lexical phonological processes have operated. In fact, what we find in CELEX is a mixture of things. In the Dutch and German databases there are some stems for which an underlying form is provided. This form is derived automatically from morphological and phonetic information, and is only provided for a small subset of the words. In the case of the Dutch databases, some of these forms are presented before final consonant devoicing and resyllabification, for instance. For example, the Dutch word *werkgeefster* ('employer') has the underlying ("phonological" transcription $wErk\#xe:v\#ster^3$ and the surface ("phonetic") transcription $[wErk] [xe:f] [st@r]$. However, these forms are only provided for morphologically complex words, with German lemmas such as *Rad* having Rat as both the underlying and surface form. In the English databases, the lemma phonological forms are simply the surface forms of the citation forms of the lemma. This is rarely different from the stem in English anyway, but in German, the phonological form of a verb, for instance, is the surface form of the infinitive form of the verb, after lexical and post-lexical processes, including resyllabification. Thus the verb *antworten* 'answer' has $'\&nt-vOr-t@n$ as its lemma phonology. So in some cases for Dutch and German, the databases do have the underlying form for the lemma representation but for all of the English and most of the other databases the phonological (phonetic/phonemic) representations are actually the surface word forms, with the lemma phonology being the surface form of the headword.

3.2 Orthography databases

The orthography databases in each language provide spelling variants. For English that involves cases where there are different spellings in different varieties, largely UK vs US spellings (*colour* vs *color*). For Dutch a distinction is made between "Preferred", "Non-preferred" (but still standard) and "Informal" (non-standard) spellings. For English, the lemma database has similar alternatives for

³ The phonological forms from CELEX are given here using SAMPA (Wells 1987).

British and American spellings. For German, there are no alternative spellings. For all three languages the underlying databases provide forms with diacritics (in ASCII format) and scripts can be used to strip these diacritics, if desired.

For the PolyOrth project, the orthographic databases in CELEX were only used to test the output of the lexicons. If the aim had been to find an engineering solution, providing orthographic forms for the three languages, the CELEX orthographic forms could have simply been incorporated into the lexicons. However, the aim was to develop a theory of how the orthographic forms could be derived from the phonological and morphological information and it was useful to be able to test the output of the theory against the existing databases.

3.3 Phonology databases

For the purposes of this paper, we are actually more interested in the phonological databases, and the issues really begin with the naming of these databases. The documentation that comes with CELEX labels the databases “Phonological”, but the descriptions mostly use the term “phonetic”. This immediately highlights a vagueness over a crucial distinction. As mentioned above, the brief description of the syllable frequency databases even mentions “phonemic” transcriptions. The PolyLex lexicons aim to provide an underlying representation, as depicted in the diagram from Nunn in figure 1. This was the assumption behind the plan to derive orthographic forms from the underlying phonological representations. We will come back to this at some length. For now, let us simply note what information is included in these databases.

The CELEX phonological databases include four distinct, but to all intents and purposes interchangeable, computer readable phonetic transcriptions. These are SAMPA (Wells 1987), CPA (Computer Phonetic Alphabet, also known as Esprit 291, developed at the University of Bochum, Germany) and two systems devised specifically for CELEX: the CELEX set and the DISC (Distinct Single Character) set. For the PolyLex lexicons the SAMPA representations were used as it is the most widely recognised.

The phonological forms are provided with syllable boundaries (marked with square brackets in the SAMPA transcriptions. The English SAMPA representation of the word *colour* is [kV][l@r*] (the rhotic “r”, only pronounced in certain accents is marked with “r*”).

Another point to note here is that, even though the orthographic databases distinguish British and American spellings, there is no such distinction made between pronunciations, even though there is undoubtedly much greater variation in the pronunciation. The primary pronunciations are taken from Daniel Jones’ *English Pronouncing Dictionary* (the edition is not specified). There are, in some

cases, large numbers of secondary pronunciations, but no systematic representation of UK/US (or any other accent) differences.

4 PolyLex and PolyOrth

4.1 PolyLex

The PolyLex project (Cahill & Gazdar 1999) aimed to explore the relationship between phonology and morphology and to what extent that relationship was consistent across related but distinct languages. The aim was to develop a lexical framework which would allow the testing of a theory and its application to the closely related languages English, German and Dutch. The morphology and phonology of these three languages share significant features, due to their historical relationship, yet the languages display significant differences in both. The languages are also well documented. This meant that a lexical framework could be built around these three languages which could ultimately be extended to cover other language families, as well as other dialects or varieties of a language.

The lexicons are defined as inheritance hierarchies in which nodes represent lexical entries or lemmas as well as word classes and sub-classes. Default (lexical) information is inherited by nodes lower in the hierarchy. For example, the noun *sheep* inherits information about its plural from a node representing the class of nouns that do not inflect for the plural. This in turn will inherit information about word class from a node representing all nouns and so on. The nodes at the bottom of these hierarchies represent individual lemmas. Each lemma node can be queried with a path that represents the morphosyntactic features of a specific form and the output to this query is a fully inflected form. Thus we can query the lexicon with the node-path pair: Walk:<past tense> and get the result [wO:kt].

The development of the PolyLex lexicons involved two key stages: initial hierarchy construction followed by automatic extension. The first part of this process was entirely manual and based on knowledge of the grammatical (specifically morphological) patterns of the three languages. The automatic extension process involved using the CELEX Databases to populate the lexicon.

Hierarchy construction

The PolyLex lexicons consist of four inheritance hierarchies: one for each language and one representing shared multilingual information. Sets of verbs, nouns, adjectives, adverbs, pronouns, articles and prepositions were chosen which represented all of the different inflectional behaviours of that language. For example, for English a hierarchy was constructed for verbs that included all of

the non-regular⁴ verbs, representing their relationships in terms of default inheritance. The aim was to manually construct the hierarchy in terms of morphological behaviour such that any new items added to the lexicon could simply be assigned to an inflectional class and inherit all information about their morphology from that class.

Automatic extension

The lexical hierarchies that had been manually constructed included a small number of example lemmas⁵ for each class, and so their coverage of the vocabularies of each language was very limited. The inclusion of the irregular and sub-regular lemmas did mean that the most frequently occurring lemmas were present, but the numbers of lemmas was still only in the hundreds. An automated process was therefore used to populate the lexicons. It was vital for this process to have access to databases that conformed to Nerbonne's definition. That is, they had to be declarative and consistent, both within and across the languages.

The CELEX Databases link the lemmas and word forms, so that it was possible to use these to determine from the word form databases every word form that belonged to a particular lemma. This was crucial in determining the inflectional class. To use a relatively simple example, for a German noun, the singular nominative form, the singular genitive form, the singular dative form and the plural nominative form would determine categorically which of the inflectional classes the lemma belonged to. This is reminiscent of Finkel & Stump's (2007) Principal Parts approach to inflectional class membership. Once we had determined which inflectional class the lemma belonged to, we could assume that the morphological behaviour would follow. For the PolyLex lexicons we then just needed to determine the phonological structure of the stem, and the realisation of the inflected forms would follow.

The definition of phonological forms in PolyLex, being founded on Syllable-Based Morphology (Cahill 1990), required the stem form to be divided into syllables, but this was provided by the forms in the CELEX databases, as was the primary lexical stress. The process to determine the lexical entries simply required recognising the vowel in each syllable and associating the consonants before it with the onset and the consonants after it with the coda. All syllable constituents were defined as being empty if undefined.

⁴ The term "non-regular" is used to cover irregular but also sub-regular verbs – i.e. those which follow a pattern but which are not part of the so-called fully regular verb class.

⁵ In the PolyLex/Orth publications and documentation the term "lexeme" is used, but the status of these lexemes is essentially the same as the lemmas in CELEX.

The automatic process therefore took the syllabified SAMPA representation from CELEX and derived from that the onset, peak, and coda of each syllable (counting from the right-hand edge) and assigned the stress to the relevant syllable. Thus, for the word *amend*, the following PolyLex lemma entry was created. The default position for stress was the final syllable, so it does not need to be specified here.

```
Amend: <> == M_Amend
      <phn syl2 peak> == @
      <phn syl1 onset> == m
      <phn syl1 peak> == E
      <phn syl1 coda> == n d.
```

In order to determine the words to be included, a list was drawn up of the 3,000 most frequent words of English. Native speakers of Dutch and German were used to provide the translation equivalents of those words in the other two languages and then any words that had been included in the manually constructed hierarchies for all three languages were added in. This resulted, therefore, in just over 3,000 words for each language, with links across the languages. The creation of the multilingual hierarchies is not relevant for this paper but is described in Cahill & Gazdar (1999). The phonological forms in PolyLex were assumed to be lexical or underlying forms, and therefore to represent the forms after lexical phonological processes but before post-lexical ones. This meant that processes like nasal place assimilation that occurred within words was included but the same process across word boundaries was not. It also meant that morphological and morphophonological processes were included, so that umlaut and ablaut processes and voicing assimilation in the English plural suffix were included. One process that we included that does not come into this category is the final consonant devoicing found in German and Dutch. We did this because we wanted to compare the output of our lexicons against the CELEX databases, to assess the accuracy of the theory implemented in our lexicons, and the CELEX database forms included final consonant devoicing. In order to do a direct comparison, therefore, we had to also include final consonant devoicing in the lexical output.

4.2 PolyOrth – orthography from phonology

The process of adding orthography to the PolyLex lexicons has been documented in detail in Cahill et al (2013). There follows below a brief outline of the process.

One of the key aims of the PolyOrth project was to determine to what extent it is possible to determine the spelling of words from their pronunciation and morphology. Specifically, the aim was to determine the number and precise nature of the phoneme-grapheme mappings, the spelling rules that are required after the initial phoneme-grapheme mapping and how many of these mappings and spelling rules are shared by the three languages. Another aim was to measure the extent of the irregularities, that is, how many of the spellings could not be predicted by the phoneme-grapheme and spelling rules. The PolyOrth lexicons made use of default inheritance, so that any cases not covered by the rules could be defined by an overriding rule. Thus the extent of the irregularity of the spellings could be determined by the number of overriding statements required to get the correct spellings.

A system of the form proposed by Nunn, as shown in figure 1, was assumed.

Slightly different approaches were taken for the three languages, according to the resources that were available for each language. For Dutch, there was an existing thorough account of the phoneme-grapheme mappings and spelling rules from Nunn (1998). It was therefore decided to adapt Nunn's rules. For German, a pilot study during the PolyLex project had made a first attempt at defining the phoneme-grapheme mappings, so that was taken that as the starting point. For English, there were a number of previous studies which provided suggested phoneme-grapheme mappings. In particular, Carney (1994) and Rollings (2004) were used as the starting point for the account. The first stage of the process involved adapting the highest levels of the hierarchies to provide two different realisation outputs: phonological and orthographic. The default information about word, morpheme and syllable structure was largely unchanged, with the assumption that orthographic syllables have the same structure as phonological ones (see Evertz & Primus (2013) for a justification of the idea of hierarchical syllable and metrical structure in orthographic forms). Thus, a syllable consists of an onset and a rhyme, regardless of whether it is a phonological or an orthographic syllable.

The next stage was to implement a set of Finite State Transducers (FSTs) to convert the phonological forms to orthographic forms. This is the key area of innovation in the PolyOrth lexicons, as the FSTs are extremely simple and each one only applies to a single syllable constituent (onset, peak, coda) within a single word. However, this does not result in massive redundancy, as the definition of the FSTs is all within the inheritance hierarchy and makes extensive use of the default inheritance mechanisms inherent within the lexicons. Thus, a set of default mappings can be defined once high in the hierarchy and specific syllables within specific morphemes (stems or affixes) can override the default. This approach to using FSTs within an inheritance hierarchy was first developed by Herring (2006).

There were four distinct default FSTs defined for each language: one for each syllable constituent to define the phoneme-grapheme mappings and one to run over the output of this process to define the spelling rules for each language. In addition, there was a further set of four FSTs which captured the shared mappings across the languages. The multilingual aspects are not of concern to us here, but the reader is directed to Cahill et al (2013) for more detail about this and other implementation details of the PolyOrth lexicons.

4.2.1 Phoneme-grapheme mappings

As mentioned above, the assumption at the start was that the phonological representations derived from CELEX would be underlying forms, which would map to orthographic forms as in Nunn's model, albeit with differing degrees of regularity in the three languages. The first stage, therefore was to determine the appropriate default mappings for each language.

Assumptions of what were the default mappings were based on statistical measures of the actual realisations. There are a number of potential issues with this approach, such as whether type or token frequency is more important, but the starting point was the set of figures or suggested mappings provided by Carney, Rollings and Nunn. Some, especially for the consonants, were obvious: /p/ in English is most commonly realised orthographically as <p>. Others, especially the vowels, were less so: /i:/ in English may be written as <ee>, <ea>, <ie> and many others. The use of default inheritance meant that we were able to overcome these irregularities by defining a single default mapping for each phoneme and overriding this at the level of the lemma or, if appropriate, word class or sub-class.⁶

Another important aspect of the PolyOrth lexicons was the use of morphological information as well as phonological to determine the spelling. For example, in English, the plural suffix is orthographically realised as <s> regardless of whether it is pronounced with /s/ or /z/. This made it possible to make use of the fact that mappings were defined at the level of the morpheme to specify the spelling of the plural suffix as independent of the pronunciation.

The spelling rules do slightly different things in the three languages. For example, in Dutch they account for the consonant degemination/vowel doubling which involves resyllabification, but in English they mostly just account for phenomena like the fact that a stem-final <y> alternates with <ie> depending on whether it has a suffix. German has very few spelling rules, but shares rules such

⁶ In theory it would be possible to set up a parallel inheritance hierarchy which grouped lemmas into different classes according to their use of different phoneme-grapheme mappings but in practice **we** only added overriding at the lemma level.

as that three identical letters in a row are reduced to two.⁷ The spelling rules are not of significance to the issues addressed in this paper but the reader is again referred to Cahill et al (2013).

5 Problems

This section addresses three specific areas that illustrate problems in the development of the lexicons from the CELEX databases. These three specific problems involve the representation of schwa vowels in English, final consonant devoicing in German and Dutch and umlaut/ablaut in all three languages.

The status of schwa in English is interesting. It clearly represents a context-determined variant, but it does not fit neatly into the role of allophonic variant. If we want to go back to the ideas of the classical phonemicists it might be viewed as an archiphoneme. The precise nature of this status is not really important here, what is important is that the presence of the schwa in the forms provided by CELEX shows us that they do not have the status of an underlying representation that can be simultaneously mapped to either spelling or surface pronunciation. In phonological terms the most accurate way to describe the status is that it occurs where the contrast between vowels has been neutralised, specifically in unstressed syllables. However, there is a real sense in which the underlying form of vowels which are realised as schwa in their normal surface realisation is **not** schwa. This is an example of the distinction between lexical and post-lexical phonology. The reduction of vowels to schwa is a post-lexical process, which is dependent on stress placement as well as factors like speech rate. As such, we would not expect these vowels to be represented as schwa in a lexicon. If we go back to Nunn's model, the abstract sounds that map either to spellings or pronunciations need to be full vowels. They need to be full vowels to determine the correct spelling, but they need to be full vowels for certain forms of pronunciation as well.

The extent of the problem in PolyOrth was different for the different languages. The key aim of the approach was to determine the relationship between the phonological, morphological and orthographic levels of representation and determine the most efficient way to represent this relationship in a set of mappings and rules. The resulting lexicons were evaluated by comparing the total number of overriding mapping statements to the total number of phonological realisation statements. As explained in Cahill et al (2013:169), this is only an approximation, but it is the closest available, relatively simple measure. According to this measure, English requires an overriding rate of 32%, higher than the other two languages, as expected. However, if all the instances with an underlying schwa are removed, this reduces significantly to 23%. Thus, nearly a third of all overriding

⁷ Following the spelling reform of 2006, it is actually allowed for a sequence of four <s>s in some varieties of German.

statements required in the English lexicons are due to the appearance of an underlying schwa rather than the phonological underlying full vowel.

Vowel alternations include various forms of ablaut including German umlaut but also ablaut in English verbs such as *swim/swam/swum*. All three languages share sets of verbs which behave in similar ways (German *schwimmen/schwam/geschwommen*, Dutch *zwemmen/zwom/gezwommen*). Consonant alternations include principally final consonant devoicing. In order to deal with these two forms of alternative realisations the PolyLex lexicons introduced an explicit notion of underlying (or lexical) and surface forms. The assumption was that any syllable constituent might have distinct underlying and surface forms, but by default the underlying and surface forms were the same. The underlying form in each case is indicated with a hyphen, so an underlying peak is labelled “peak-“, and underlying coda “coda-“ etc.

Each stem had an underlying vowel which was marked by using the label “peak-“. By default, any surface peak was defined as being the same as the underlying peak, but in forms where an ablaut operated the surface form was defined as distinct from the underlying form. Exactly how the surface form was defined varied (umlaut was defined as a single operation which had pre-defined mappings from underlying to surface forms, but ablaut in English verbs was defined at the verb class level), but the relationship between the underlying and surface form was the same. These “surface” forms are the forms after lexical phonological processes, but not after post-lexical processes.

The consonant alternations accounted for were cases of what is generally termed “final consonant devoicing”. This occurs in both Dutch and German standardly as a post-lexical phonological process. All voiced obstruent consonants (i.e. plosives and fricatives) are devoiced when they occur in a syllable coda. The same way of defining underlying and surface forms was used for this as for ablaut, with “coda-“ being the underlying form and “coda” the surface form of a final consonant cluster, with an equation stating that they were, by default, identical.

It was always clear that the consonant devoicing was a different kind of process from the ablaut processes. Specifically, it was clearly post-lexical, while the ablauts were lexical and, indeed, morphological rather than phonological. However, for the PolyLex lexicons, these approaches to vowel and consonant alternations allowed **us** to evaluate the output of the lexicons by means of direct comparison with the CELEX databases. The lexicons were defined and their output evaluated against the forms provided by CELEX. The two lemmas *Rat* and *Rad*, for example, had the same phonological representation in CELEX but the word forms *Räte* and *Räde* have the different consonants reflecting their different surface pronunciation. These word forms actually represent the distinct underlying forms, but this is not reflected in the lemma phonology. It was only when

orthography was added in the PolyOrth project that the problem with this approach became clear. The problem stems from the fact that there is a complex relationship between phonological underlying and surface forms, lemmas and word forms and most importantly between these and orthography.

Although not entirely consistent, it is generally the case that post-lexical phonology is not reflected in the orthography whereas morphology and morphophonology/lexical phonology is. These tendencies are reflected fairly consistently in the two examples above. In German, English and Dutch, ablaut processes are reflected in the orthography, with different vowels in the spelling as well as in the pronunciation, while final consonant devoicing in German and Dutch, which is a post-lexical process, is not. Thus *Rad* and *Räder* (*wheel/wheels*) both have a <d> in their spelling, even though they are pronounced /rɑ:t/ and /re:də/. Note that the vowel umlaut is reflected in the orthography here.

This distinction means that the status of the relationship of the spelling to the “peak-” and to the “coda-” are different. The orthography should refer to the **underlying** form of the coda, before devoicing has occurred, but it should refer to the **surface** form of the peak, after umlaut or ablaut has occurred. This is not necessarily a problem, as the phoneme-grapheme mappings can simply be defined as involving the surface “peak”, but the underlying “coda-”. It is, nevertheless, unfortunate that what appears to be uniform notation is indicating different things. However, the problem with codas is actually more difficult than this. Interestingly, there is also a similar voiced/voiceless relationship in English, in a handful of words such as *knife/knives*. This is a case of a complex synchronic/diachronic relationship. In this case, in fact, the historical alternation involved voiceless fricatives becoming voiced when they occur between two voiced sounds. There is a crucial distinction in the phonological level at which the voicing alternations operate. In Dutch and German the underlying form of words like *Rad* have a voiced obstruent, which is reflected in the spelling. In English, the underlying representations of singular and plural forms such as *knife/knives* have different consonants. This is not a synchronic post-lexical process in English, and this difference is reflected in the spelling. In CELEX, the phonological forms for these English words have the correct underlying consonants, but the Dutch and German have the surface forms. Thus, in relation to Nunn’s model, the English databases have the abstract sounds required but the Dutch and German do not.

The CELEX databases make a clear morphological distinction between lemmas and word forms. This would appear to provide one way in which a distinction between underlying and surface forms could have been represented. In the case of the example of final consonant devoicing above, the surface

word forms *Rad* [Rat] and *Räde* [Re:d@] rightly have distinct consonants, as they are both surface word forms. However, the phonological form of the lemma should not reflect the surface realisation, after post-lexical phonological processes have operated. Thus the CELEX lemma databases should be making the distinction between the underlying and surface forms in these cases, as they do for a relatively small number of stems.

The situation in Dutch is slightly more complex, as discussed in both Nunn (1998:54-55) and Sproat (2012:37-38). For most obstruents, the relationship is as discussed above. The spelling reflects the voiced, underlying consonant. However, in the cases of /f/ and /v/ and of /s/ and /z/, the surface forms are reflected in the spelling. Thus the verb *lezen* ‘to read’, has forms *ik lees* ‘I read’ and *wir lezen* ‘we read’. Nunn accounts for these examples by the use of an autonomous spelling rule, termed “orthographic devoicing”. The phoneme-grapheme mappings access the pre-devoicing underlying form, but these two consonants are respelled to reflect their devoicing when they occur in the coda of an orthographic syllable. Sproat further points out that the devoicing of the past tense suffix, involving the consonants /t/ and /d/ is consistently represented in the spelling. Thus, there are verb forms *kamde* ‘combed’ and *hoopte* ‘heaped’. Sproat explains this apparent discrepancy by stating that there are two different types of devoicing operating in Dutch. The first, applying to any coda obstruent, is post-lexical and is not represented in the mapping from phoneme to grapheme. The second is a case of voicing spreading from a preceding consonant. Sproat’s argument is convincing, but it nevertheless reveals that determining exactly what the appropriate level of representation is can be extremely complex and requires a call to subtle (and potentially controversial) linguistic analyses. If we compare the Dutch example to English, where voicing assimilation in both the <s> and <ed> suffixes is not reflected in the spelling, it is clear that there are some very fine-grained distinctions to be made in determining the ORL for each language.

In the case of ablaut, the representations are as expected. The different vowels in the different forms of *swim* only occur in the word form database. The lemma database only has the citation form with the vowel /ɪ/. This clearly points to the fact that the representation of the lemma is of a form before morphology has operated, let alone morphophonology (lexical phonology) or post-lexical phonology. However, the lemma/word form relationship in CELEX is actually a relationship between a set of different surface forms, one of which has special status as the headword or citation form.

For the purposes of testing theories about the relationship between phonemes and graphemes, it is important to have an appropriate **phonemic** level of representation.

6 Discussion

The issues discussed above raise questions about exactly what is necessary or useful in lexical databases for the purposes of research on writing systems.

One of the fundamental problems with databases such as CELEX is that, unlike for the orthographic representations, there is no single definitive phonological/phonetic representation. Phonetic representations are, by definition, variable at the level of individual utterance, let alone at the level of individual speaker or accent. Phonological representations are abstract and depend on theory specific criteria that are not agreed by phonologists. What the CELEX databases attempt to do is give a reasonable representation of an average pronunciation of each word, much in the way that a dictionary gives **us** a representation of the pronunciation of a word in isolation, in careful speech by a speaker of the standard form of the language. This necessarily abstracts away from issues such as context (position in sentence, influence of prosodic and intonational patterns etc.), speaker's accent, speed of speech, mood or attitude of the speaker, age or gender of the speaker etc.. There is, however, no definitive way of representing this kind of abstraction – it is a level of representation that is bound to be subjective, at least to some degree.

In terms of theoretical linguistics, there is the well-established concept of the phoneme, which might be seen as the answer to our problems. However, there is no universally agreed definition of the phoneme which could be used for a database like this. The strict phonemic principle, as defined by the original phonemicists (e.g. Swadesh, 1934), has long been discredited and in part because of examples like the schwa problem discussed above. A phonemic representation of the word *construct* would be the same for the noun and the verb: /kɒnstʃʌkt/. There would then be rules which determined the stress pattern for the noun and the verb and reduced the vowel in the unstressed syllable appropriately to get the actual phonetic realisation.

One key question that must be asked is what the purpose of the database is. The CELEX databases were developed as multi-purpose repositories of information and, as such, attempts were made to include a variety of different levels of information which might be useful for different purposes. This applies to all of the different databases. The orthographic databases, for example, include representations with and without diacritics. While this has little impact on the English forms, it means that there are rather different forms for German depending on whether umlaut and eszett are used and similarly for the diaeresis in Dutch. For the phonological databases the different versions of computer readable phonetic alphabets used have already been mentioned. The databases also include CV-template and syllabified representations. For the specific purposes of the PolyLex/Orth lexicons, it would have been helpful to also have more genuinely abstract, underlying “phonemic”

representations. But is this a reasonable thing to expect? What possible other functions could such a representation serve?

The developers of CELEX are clear about the fact that the genuinely underlying phonological forms are only provided for a minority of morphologically complex words and also explain that this is because the process of deciding on such forms is not trivial and requires knowledge about morphology, phonology (including the nature of post-lexical processes in the language in question) and orthography. The developers of the LEXIQUE and CLEARPOND databases have manually attempted to recover an underlying phonological level from raw phonetic input. However, this process is unlikely to be fully consistent and is dependent on an interpretation of what that level should be that is not agreed by linguists.

This brings us to the fundamental question: what are the “phonemes” in phoneme-grapheme (or grapheme-phoneme) mappings? In the field of computational speech processing, this is not an issue, provided there is a level of representation on which researchers are agreed. Developing computational tools that work with that level of representation is the key aim and there is no need to worry about exactly what the linguistic status of those representations is. However, for researchers interested in questions of the theory of these relationships, it is vital that we are clear that different theories of phoneme-grapheme relationships are actually assuming the same “phonemes”. It is notable that Nunn does not use the term “phoneme”, preferring the term “abstract sounds” (although she goes on to say that she will refer to the mappings between abstract sounds and spellings as phoneme-grapheme mappings). Nunn discusses several issues relating to the status of the abstract sounds, albeit specifically in relation to Dutch. The status of schwa is of particular interest. Nunn states that schwa in Dutch does not represent an abstract vowel sound at all. This is clearly not the case for schwa in English. Another feature of Nunn’s approach is that she proposes that phonological and orthographic syllables are not necessarily identical. She justifies this with examples such as *zingen*, [zɪN@]. The two graphemes <ng> which together represent the single underlying sound /N/ lead to an interpretation with two orthographic syllables, *zin.gen*, and two phonological syllables, /zɪ.N@/. The CELEX treatment of this word reveals a similar distinction between orthographic and phonological syllable boundaries, with the orthographic syllabification as in Nunn, and two different phonological syllabifications: 'zɪ-N@ and [zɪ[N]@]. In the English CELEX, the word *banger* has the orthographic syllabification *bang.er*, and the phonological syllabification 'b{-N@R. The fact that most phonologists would not suggest this syllabification due to the impossibility of the /N/ occurring in an onset suggest that the same care over the status of orthographic and phonological syllabification does not appear to have been taken in the English databases.

Alphabetic writing systems are generally assumed to be, in some sense, phonemic. Thus, a language like Spanish has a spelling system in which each phoneme of the language is paired with a grapheme and this set of mappings allows us to predict the pronunciation from the spelling and vice versa. However, as discussed by Sproat (2000), different writing systems are based on different levels of the phonology. Such distinctions make it tricky to determine what the ideal level of representation of abstract phonology should be.

The relationship presented above in Nunn's figure represents the ideal but, as we have seen, it is certainly not representative of the forms included in the CELEX phonological databases. The PolyOrth project was an attempt to determine the relationship between the phonology, morphology and orthography of the three languages, but its conclusions are only valid up to a point, with the proviso that what is meant by "phonology" here is somewhat unclearly defined. In order to fully investigate the relationship, it will be necessary to recreate the PolyLex lexicons with genuinely abstract (underlying) phonological forms, with a principled distinction between lexical and post-lexical phonology. However, identifying exactly what these forms should be would be a significant research project in its own right. Most importantly, there is a conflict between what is required from a database providing phonological forms to suit the orthographic depth of different writing systems and the idea of consistency of representation across languages. Ultimately, in order to analyse the various relationships across languages, we need databases with both deep or underlying forms and surface forms. We can then determine whether each language/writing system uses the underlying or surface phonology as the basis for its orthography and indeed test Sproat's claim that the relationship is consistent within a single writing system.

7 Conclusions

The existence of lexical databases for various levels of linguistic information is of potential use for a range of applications. For computational work, such databases can provide ready-made lists of vocabulary that can be used more or less directly by NLP systems. For linguists, they can be used to test theories on large numbers of possible examples. However, while orthographic databases can be relatively definitive, due to the standardised nature of written language, phonological databases depend on a range of decisions about the level of abstraction of the representations included. Depending on the level chosen, there are also questions about things like the level of formality and the precise social or regional accent represented. For the specific purpose of researching writing systems, lexical databases of both orthographic and phonological forms would appear to be a useful source of data. However, the examples discussed above show that the development of large databases that include genuinely phonological underlying forms is not a simple process. Decisions

must be made about the exact level of representation, and those decisions may not be appropriate for all researchers. Even if an ORL as proposed by Sproat is agreed upon, this level will be different for different languages and writing systems and will thus make cross linguistic comparison difficult. It also raises questions of circularity. Depending on what is orthographically relevant for each language does not provide *a priori* neutral phonological representations to map to spellings. However, the process of developing databases of the ORL for each language would require researchers to determine what that level was in each case which would be a useful and revealing process in itself.

References

- Baayen, Harald, Richard Piepenbrock & Hedderik van Rijn (1995). *The CELEX lexical database, Release 2* (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.
- Benesty, Jacob, Sondhi, M. M., Huang, Yiteng (Eds.) (2008). *Springer Handbook of Speech Processing*, Berlin: Springer.
- Brown, Dunstan & Andrew Hippisley (2012). *Network Morphology: A Defaults-based Theory of Word Structure*, Cambridge: CUP.
- Burnage, Gavin (1996). *CELEX: A Guide for Users, The CELEX lexical database, Release 2* (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania..
- Cahill, Lynne (2001). Semi-automatic construction of multilingual lexicons. *Machine Translation Review*. Electronic journal, available at www.bsc.org.uk/siggroup/nalantran/mtreview/mtr-12/.
- Cahill, Lynne (1990). Syllable-based Morphology. *Proceedings of the 13th International Conference on Computational Linguistics (COLING90)*, Vol. 3, Helsinki, Finland, August 1990, 48-53.
- Cahill, Lynne & Gerald Gazdar (1999). The PolyLex Architecture: Multilingual Lexicons for Related Languages. *Traitement Automatique des Langues* 40.2: 5-23.
- Cahill, Lynne Carole Tiberius & Jon Herring (2013). PolyOrth: Orthography, phonology and morphology in inheritance lexicons. *Written Language and Literacy* 16.2: 146-185.
- Carney, Edward (1994). *A survey of English Spelling*. London: Arnold.
- Carroll, J. & C. Grover (1989). The derivation of a large computational lexicon of English from LDOCE. In B. Boguraev & E. Briscoe (eds.) *Computational Lexicography for Natural Language Processing*, 117-134. Harlow, UK: Longman.

- Evans, R., P. Piwek, L. Cahill & N. Tipper (2008). Natural Language Processing in CLIME – a multilingual legal advisory system, *Journal of Natural Language Engineering* 14:1: 101-132.
- Evertz, Martin & Beatrice Primus (2013). The graphematic foot in English and German. *Writing Systems Research*, 5.1: 1-23. doi: 10.1080/17586801.2013.765356
- Finkel, Raphael & Gregory Stump (2007). Principal Parts and Morphological Typology. *Morphology* 17.1: 39-75.
- Goldrick, Matthew & Brenda Rapp (2007). Lexical and post-lexical phonological representations in spoken production, *Cognition* 102: 219-260.
- Herring, Jon (2006). *Orthography and the lexicon*. PhD Dissertation, University of Brighton.
- Nerbonne, John (1998). *Linguistic Databases*. CSLI (ISBN: 9781575860930)
- New, Boris, Christophe Pallier, Marc Brysbaert & Dominic Ferrand (2004). Lexique 2: A New French Lexical Database, *Behavior Research Methods, Instruments, & Computers* 36: 516. doi:10.3758/BF03195598
- Nunn, Anneke (1998). *Dutch Orthography: A systematic investigation of the spelling of Dutch words*. The Hague: Holland Academic Graphics.
- Polich, John, Gregory McCarthy, William S. Wang & Emanuel Donchin (1983). When words collide: Orthographic and phonological interference in word processing, *Biological Psychology* 16.3-4: 155-180
- Rollings, Andrew G. (2004). *The spelling patterns of English*. Munich: Lincom.
- Sampson, Geoffrey (1985). *Writing Systems: A Linguistic Introduction*. Stanford: Stanford University Press.
- Sprenger-Charolles, Liliane, Linda S. Siegel & Danielle Bechennec (1998). Phonological Mediation and Semantic and Orthographic Factors in Silent Reading in French, *Scientific Studies of Reading* 2.1: 3-29.
- Sproat, Richard (2012). The Consistency of the Orthographically Relevant Level in Dutch, in Neef, Martin, Anneke Neijt & Richard Sproat (Eds) *The Relation of Writing to Spoken Language*, 35-46. Berlin, Boston: Max Niemeyer Verlag.
- Sproat, Richard (2000). *A computational theory of writing systems*. Cambridge: CUP.
- Swadesh, Morris (1934). The Phonemic Principle. *Language* 10.2: 117-129.

Wells, John C. (1987). Computer coded phonetic transcription. *Journal of the International Phonetic Association*. 17.2: 94-114.